Appendix 2 (as supplied by the authors): Supplemental material

Keyword searching algorithm development:

We initially came up with search terms for each of the 23 diagnostic groupings based on clinical expertise of the investigators and then refined it in an iterative manner. Additional search terms, common abbreviations, misspellings, and idiosyncratic descriptors were included after manual review of a subset of notes that were accompanied with a relevant modified ICD-9 code. We created frequency tables of terms that appeared in proximity to our keywords to create a list of negations/exclusion terms to refine our search strategy. Lastly, we compared encounters that were identified by keywords or bills for particular disease groupings to examine discrepancies and refine the search strategy.

Validation analysis:

To evaluate the impact of the accuracy of diagnostic definitions using administrated electronic medical record data from the Electronic Medical Records Primary Care database (EMRPC; also known as EMRALD) the primary author performed a manual chart review of 96 records. The charts were stratified, evenly distributed, and randomly sampled by condition. The reviewer was blinded to the diagnosis and by reviewing only the physicians' typed notes, we assigned the most likely diagnosis based on the physician impression. This was used as the gold standard and compared to 6 different diagnosis definitions. If multiple diagnoses were present a hierarchy was utilized that assigned the diagnosis to be the conditions with the highest antibiotic appropriateness rate (for example pneumonia is higher than common cold).

Diagnosis sensitivity analysis:

In Ontario, physicians can only bill 1 diagnosis per patient per day. It is possible patients presented with multiple complaints resulting in diagnostic misclassification. To evaluate the impact of different diagnosis algorithms that differentially utilized keyword searching over billing claims we performed a series of sensitivity analyses. The definitions evaluated were: (1) billing claim only using modified ICD-9 codes, (2) keyword search only from the typed progress notes, (3) billing code then if absent to search remaining records for a relevant keyword (primary analysis), (4) keyword search then if absent to search remaining records for a relevant billing code, (5) search all records for either a relevant billing code or a keyword, (6) and finally we used the primary analysis but reversed the diagnostic hierarchy when more than one diagnostic billing code or keyword was identified. This reverse hierarchy assigned the diagnosis to the condition with the lowest antibiotic appropriateness rate (for example common cold was higher than pneumonia).

We determined the percent accurate diagnosis for each definition. Then using each of the definitions we repeated the analysis for the entire cohort to calculate the overall unnecessary antibiotic prescribing rates using the same methodology as the main manuscript. We then categorized the errors as having no clinical significance, potentially overestimating the appropriateness rate, or potentially underestimating the appropriateness rate.

The results are summarized in Figure S1 and S2. The y-axis is the percent accuracy of diagnosis using various definitions from manual chart review. Each bar represents one of the 6 diagnosis definitions. Note that n=96, except for n=63 for the billing only definition. The different definitions resulted in minor changes in accuracy. Billing codes only had the lowest accuracy at 60% and keywords had the highest at

71%, and the accuracy of the primary analysis was 67%. Of the errors in the primary analysis, approximately 1/3rd had no clinical significance (for example asthma diagnosis misidentified as common cold, which have identical expected antibiotic appropriateness rates). Approximately 2/3rd of the errors resulted in a condition with a higher expected antibiotic prescribing rate (for example the true diagnosis was common cold but classified as pneumonia for this study). This would result in potentially underestimating the unnecessary antibiotic prescribing rate. The different definitions resulted in relatively small changes to the primary outcome of antibiotic inappropriate prescribing rates. When we applied each of these 6 definitions to the entire cohort, the antibiotic inappropriate prescribing rates varied from 13% to 18%, with a rate of 15% for the primary analysis (Figure S2).

This validation analysis demonstrated that all definitions have some associated error. However, there was no single optimal definition but the incorporation of keyword searching to physician billing claims did improve accuracy. The clinical significance of this error is likely to be small either by having no impact or by making the estimates of unnecessary prescribing conservative. We also demonstrated that the validity of our primary analysis was robust to multiple different definition assumptions.

Appendix to: Schwartz KL, Langford BJ, Daneman N, et al. Unnecessary antibiotic prescribing in a Canadian primary care setting: a descriptive analysis using routinely collected electronic medical record data. *CMAJ Open* 2020. DOI:10.9778/cmajo.20190175.